Reporting the Results of Your Study:

A User-Friendly Guide for Evaluators of Educational Programs and Practices

October 2005

This publication was produced by the Coalition for Evidence-Based Policy, in partnership with the What Works Clearinghouse, under a contract with the U.S. Education Department's Institute of Education Sciences (Contract #ED-02-CO-0022). The views expressed herein do not necessarily reflect the views of the Institute of Education Sciences.

Purpose and Overview of this Guide

Purpose: To provide clear, practical advice on reporting the results of an evaluation of an educational program or practice ("intervention").

Specifically, this is a guide for researchers, and those who sponsor and use research, to reporting the results of "impact" studies – that is, studies which evaluate the effectiveness of an intervention by comparing participants' educational outcomes (e.g., reading or math skills) with:

(i) those of a control or comparison group that does not receive the intervention, or
(ii) participants' pre-intervention ratings on these outcome measures.

The Guide suggests key items to include in the study report so as to give the reader a clear understanding of what was evaluated, how it was evaluated, and what the evaluation found.[1]

Overview: The following is an outline of the Guide, showing the key items we suggest you include in each of the sections that typically comprise a study report.

Section of the Study Report:	***Key Items To Include:***
1. Title and Abstract	A. A clear, informative title. B. A "structured abstract," including identification of the research design.
2. Background and Purpose	A. Background information on the intervention being studied. B. Purpose of the study, including the research question(s) it seeks to answer.
3. Methods	A. Description of the study setting (e.g., place and time it was conducted). B. Description of the study sample (including number of sample members and how they were recruited into the study). C. Concrete details of the intervention, and how it differed from what the control/comparison group received. D. Description of how the study sample was allocated to intervention and control/comparison groups. E. Description of how and when outcomes were measured (including evidence that the tests/instruments used to measure are reliable and valid). F. Statistical methods used to compare outcomes for the intervention and control/comparison groups (or outcomes before and after the intervention).
4. Results	A. Indicators of whether the study was successfully carried out (e.g., amount of sample attrition). B. Any descriptive data on how the intervention was actually delivered in the study (e.g., extent to which participants completed the intervention). C. Estimates of the intervention's effect on all outcomes measured. D. Any estimates of its effect on subgroups within the study sample. E. If analyzed, any estimates of its effect on those who received greater versus lower "doses" of the intervention.
5. Discussion	A. Interpretation: what the results say about the intervention's effectiveness. B. Extent to which the results may be generalizable to others who receive or could receive the intervention. C. Significance of the results to educators, policymakers, and researchers. D. Factors that may account for the intervention's effect (or lack thereof). E. Any study limitations (e.g., small study sample, sample attrition).

1. Key items to include in the Title and Abstract section

A. A clear, informative title (Illustrative example: "Randomized Controlled Trial of a Peer-Tutoring Program for Second Graders: Effect on Math Achievement Two Years Later").

B. An abstract of the study (1-2 pages), which:

- **Should follow the "structured abstract" format suggested by the U.S. Education Department's Institute of Education Sciences** (see appendix for an example).[2]

- **Should identify the type of research design used in the study.** Common examples include:

  - **Randomized controlled trial** – a study that measures an intervention's effect by (i) randomly assigning individuals (or other units, such as classrooms or schools) to a group that participates in the intervention, or to a control group that does not; and then (ii) comparing outcomes for the two groups.

  - **Comparison-group study (also known as a "quasi-experimental" study)** – a study that measures an intervention's effect by comparing outcomes for intervention participants with outcomes for a comparison group, chosen through methods other than random assignment. We suggest you also indicate if the comparison-group study is one of the following two types, generally regarded as among the stronger comparison-group designs:

    - A comparison-group study with equating – a study in which statistical controls and/or matching techniques are used to make the intervention and comparison groups similar in their pre-intervention characteristics.

    - A regression-discontinuity study – a study in which individuals are assigned to intervention or comparison groups solely on the basis of a "cutoff" score on a pre-intervention measure (e.g., students scoring at or below the 25th percentile on the Iowa Test of Basic Skills in math are assigned to the intervention group, and those scoring above the 25th percentile are assigned to the comparison group).

  - **Pre-post study –** a study that examines whether intervention participants are better or worse off after the intervention than before, and then associates any such improvement or deterioration with the intervention. [Note: The What Works Clearinghouse does not consider this type of study to be capable of generating valid evidence about an intervention's effect. This is because it does not use a control or comparison group, and so cannot answer whether the participants' improvement or deterioration would have occurred anyway, even without the intervention.]

2. Key items to include in the Background and Purpose section

A. Background information on the intervention being studied, including such items as:

- **A brief description of the intervention**, including the problem it seeks to address and the population for which it is intended. (This would be a general description, with the concrete details provided in 3C, below.)
- **The theory or logic of how it is supposed to improve educational outcomes.**
- **The intervention's history and the extent of its current use.**
- **A summary of the results of any previous rigorous studies of the intervention or closely-related interventions.**

B. Purpose of the study, including:

- **A clear statement of the research question(s) it seeks to answer** (e.g., "Did the XYZ reading program for first graders increase student reading achievement, and reduce the number of students retained in-grade or placed in special education classes? If so, by how much?").
- **An explanation of why the study is, or should be, important to educators and/or policymakers.**

3. Key items to include in the Methods section

A. A concise description of the study setting (e.g., five public elementary schools in central Philadelphia, during the period 2001-2004).

B. A description of the study population (i.e., "sample"), including:

- **How they were recruited into the study**, including (i) the eligibility requirements for participation in the study, and (ii) how those eligible were invited or selected to join the study sample (e.g., the principal of Monroe Middle School requested that all teachers in the school participate in the study of a professional development program or, alternatively, asked for volunteers to participate).[3]
- **If the sample is intended to be representative of a larger group (e.g., a representative sample of schools participating in a national program), what methods were used to obtain such a sample (e.g., random sampling).**
- **The total number of sample members allocated – through random assignment or other means – to (i) the intervention group, and (ii) the control/comparison group at the start of the study** (or, in a pre-post study, the total number of sample members prior to the intervention). You may also wish to report the results of an analysis showing that the study sample is large enough to provide meaningful answers to the study's research questions ("power analysis").[4]

- **Descriptive statistics on the study sample** (e.g., age, race, gender, family income, and pre-intervention measures of the outcomes the intervention seeks to improve, such as reading or math achievement). You should briefly describe how and when these descriptive data were obtained, and the percentage of sample members from whom they were obtained. You may also wish to discuss the extent to which the sample is typical of the larger population that receives or could potentially receive the intervention.

C. A clear description of the intervention as it was implemented in the study, and how it differed from what the control/comparison group received.

- **The description should include the concrete details of the intervention** that a reader seeking to replicate it would need to understand including, among other things, who administered it, what training or supervision they received, what it costs to implement in a typical school or community setting,[5] and how it may differ from the model program on which it is based (e.g., the National ABC Science curriculum was used but only lessons 1- 20 out of 40).

- **You should also describe clearly how the intervention differed from what the control/comparison group received.**

- **Illustrative example:** A study of a one-on-one tutoring program for beginning readers should describe such items as –

  - Who conducted the tutoring (e.g., certified public school teachers, paraprofessionals, or undergraduate volunteers);
  - The training they received in how to tutor;
  - The curriculum and materials they used to tutor;
  - The duration of the tutoring sessions, and setting in which they took place (e.g., daily 20-minute sessions held after school, over a period of six-months);
  - Whether the tutors were supervised or monitored and, if so, how;
  - Any unusual events that substantially affected delivery of the tutoring program (e.g., school closing for two weeks due to a major snowstorm);
  - The cost of the tutoring program per student (excluding costs of research or program development that would not be incurred in replicating the program);
  - The reading instruction or other services received by the students in the control/comparison group (e.g., the school's usual reading program); and
  - Where the reader can obtain additional information on the tutoring program (e.g., a website or program manual).

D. A description of how the study sample was allocated to intervention and control/comparison groups, including:

- **Whether the study allocated *individuals* (e.g., students), or *clusters* of individuals (e.g., classrooms or schools), to the intervention and control/comparison groups.** If the study allocated clusters, you should also describe any steps taken to ensure that the placement of individuals within the clusters (e.g., placement of students in classes) was unaffected by whether the clusters were in the intervention or control/comparison group (e.g., students

were assigned to classes *prior to* the allocation of classes to the intervention and control/comparison groups).

- **Whether the random assignment or formation of comparison groups was "blocked"** (e.g., students were identified as (i) high-achieving, (ii) average-achieving, or (iii) low-achieving based on prior test scores, and then allocated *within each of the three achievement levels* to the intervention and control/comparison groups).

- **Whether the ratio of those allocated to the intervention versus control/comparison group (e.g., 60:40) was held constant over time, and within blocks.**

- **Whether all those originally allocated to the intervention and control/comparison groups were retained in their group for the duration of the study – even:**

  - Intervention participants who failed to participate in or complete the intervention (retaining them in the intervention group is known as an "intention-to-treat" approach).

  - Control/comparison group members who may have participated in or benefited from the intervention (these are known as "cross-overs" or "contaminated" members of the control/comparison group).

- **In a randomized controlled trial, the random assignment process used (e.g., coin toss, lottery, or computer program), including:**

  - Who administered it (e.g., the researchers or school staff); and

  - Any steps taken to protect against intentional or unintentional manipulation of the process (e.g., concealing from researchers, school staff, and study sample members any information they could use to predict in advance who would be assigned to the intervention versus control group).

- **In a comparison-group study, how the comparison group was formed** (e.g., from students in a neighboring school who had achievement levels and demographic characteristics similar to intervention participants). Among other things, this description should include:

  - Whether the comparison group was formed before or after the intervention was administered;

  - Any "matching" techniques used to increase the initial similarity of intervention and comparison group members in their observable characteristics (e.g., propensity score matching); and

  - In a regression-discontinuity study (described in 1B, above), what cutoff score was used to form the intervention and comparison groups, how sample members' scores were distributed around this cutoff score, and any evidence that the cutoff score was used with few or no exceptions to determine who received the intervention.

E. A clear description of how and when outcomes were measured, including:

- **What tests or other instruments were used to measure outcomes** (e.g., widely-used standardized tests, such as the Stanford Achievement Test; tests designed by the research

team for purposes of the study; structured interviews with study sample members; observations of classroom behavior; school records).

- **Any evidence that these instruments:**
  - **Are "reliable"** (i.e., yield similar responses in re-tests or with different raters);
  - **Are "valid"** (i.e., accurately measure the true outcomes that the intervention is designed to affect); and
  - **Were applied in the same way to the intervention and control/comparison groups** (e.g., the same test of reading skills was administered in comparable settings).
- **Who administered these instruments, and whether any of the following conditions applied which might affect the objectivity of their measurements:**
  - They knew the study sample members (e.g., were their teachers);
  - They might have a stake in the study outcome (e.g., were the developers or implementers of the intervention being studied); and/or
  - They were kept unaware of who was in intervention versus control/comparison group (i.e., were "blinded").
- **When the instruments were administered** (e.g., just prior to the start of the intervention, so as to obtain "baseline" data"; and at 12-month intervals thereafter for three years, so as to obtain outcome data at the one-year, two-year, and three-year follow-ups).

F. The statistical methods used to compare outcomes for the intervention and control/comparison groups (or, in a pre-post study, outcomes before and after the intervention), including:

- **The regression or other model used to estimate the intervention's effects**, including – in a regression – the variables that indicate whether a sample member is in the intervention or control/comparison group, and any covariates used.
- **Any techniques used to adjust statistically for initial differences between the intervention and control/comparison groups**, or improve the precision of the estimated effects (e.g., analysis of covariance).
- **If the study randomly assigned or compared clusters (e.g., classrooms), rather than individuals, whether the study accounted for such clustering in estimating statistical significance levels.**
- **If the random assignment or formation of comparison groups was "blocked," whether the study accounted for such blocking in estimating statistical significance levels.**
- **If the study sample is intended to be representative of a larger group (e.g., a representative sample of schools participating in a national program), whether the study accounted for this in estimating statistical significance levels.** (i.e., used a "random effects" model).

4. Key items to include in the Results section

A. Indicators of whether the study was successfully carried out, including:

- **The results of statistical tests assessing the extent to which the intervention and control/comparison groups were equivalent in key characteristics prior to the intervention –** especially whether they were equivalent in pre-intervention measures of the outcomes the intervention seeks to improve, such as reading or math achievement. (Exception: regression-discontinuity studies – described in 1B above – do not seek to create equivalent intervention and comparison groups.)

- **The percentage of individuals originally allocated to (i) the intervention group, and (ii) the control/comparison group, for whom outcome data could not be obtained** at each follow-up point (i.e., the degree of "sample attrition").

- **An analysis of whether sample attrition created differences between the intervention and control/comparison groups** (e.g., whether a greater number of students with low initial test scores were lost from the intervention group as opposed to the control/comparison group, creating a difference between the two groups). You should also preferably include an analysis of whether (i) the original study sample and (ii) those for whom outcome were obtained, differ in their pre-intervention characteristics, so as to gauge whether attrition altered the nature of the study sample.

- **The extent to which any control/comparison group members participated in the intervention**, or otherwise benefited from it (e.g., by borrowing techniques or materials from intervention group members).

B. If available, descriptive data on how the intervention was actually delivered in the study, including such items as:

- **The extent to which the intervention group members completed the intervention, and what intervention services or treatment they actually received** (e.g., in a study of a teacher professional development program, the average number of training sessions the teachers participated in, the length of each session, and the extent to which the trainers covered the key items in the training curriculum).

- **Data on any related (non-intervention) services or treatment provided to the intervention group and/or control/comparison groups** (e.g., participation in other professional development activities).

C. Estimates of the intervention's effect on all outcomes measured (not just those for which there are positive effects), including:

- **Whether the effects are statistically significant** at conventional levels (usually the .05 level); and

- **The magnitude of the effects, reported in "real-world" terms that enable the reader to gauge their practical importance** (e.g., report an improvement in reading comprehension of a half grade-level, or a reduction in the percentage of students using illicit drugs from 20 to 14 percent, rather than only reporting items such as "standardized effect sizes" or "odds ratios").

- **The sample size used in making each estimate, and the standard error of the estimate.**

D. Any estimates of the intervention's effects on subgroups within the study sample, including:

- **A precise description of each subgroup** (e.g., first-grade male students scoring in the highest 25th percentile on teacher-rated aggressive behavior).
- **A brief rationale for why the subgroup was chosen to be analyzed** (e.g., reasons why the intervention might have a different effect on the subgroup than on the overall study population).
- **A complete listing of all subgroups analyzed** (not just those for which there are positive effects), and the effects found for each.
- **The sample size used in making each estimate, and the standard error of the estimate.**

E. If analyzed, any estimates of the intervention's effect on those who received greater versus lower "doses" of the intervention (e.g., those who successfully completed all sessions of a teacher training program versus those who completed few or no sessions). [Note: Such "dose-effect" analyses can sometimes yield useful information to supplement the results for the full sample. However, if intervention participants *self-select* themselves into the higher and lower dose groups, the What Works Clearinghouse would not regard such analyses as producing valid estimates of the intervention's effect (because self-selection would plausibly create differences between the two groups in motivation levels and other characteristics, leading to inaccurate results).]

5. Key items to include in the Discussion section

A. Interpretation of the study results: what they say about the effectiveness of the intervention in the context of other rigorous studies of this or related interventions.

B. The extent to which the results may be generalizable to others who receive or could potentially receive the intervention.

C. Significance of the results to educators, policymakers, researchers, and others.

D. Factors that may account for the intervention's effect (or lack thereof).

E. Any study limitations (e.g., small study sample, sample attrition).

Appendix: Example of a "Structured Abstract" using the format suggested by the Institute of Education Sciences

Abstract

Citation: Ricciuti, A.E., R.G. St.Pierre, W. Lee, A. Parsad, and T. Rimdzius. Third National Even Start Evaluation: Follow-Up Findings From the Experimental Design Study. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Washington, D.C., 2004.

Background: The Even Start Family Literacy Program has provided instructional services to low-income children and their parents since 1989. A previous randomized controlled trial in the early 1990s did not show this program to have positive impacts.

Purpose: To assess the effectiveness of Even Start in a group of grantees around the country. An earlier report from this study presented impact findings based on pretest and posttest data at the start and end of a school year. No program impacts were found. The purpose of the current report is to present impact analyses of follow-up data collected one year after posttest data.

Setting: 18 Even Start grantees in 14 states that operated in the 1999-2000 and 2000-2001 school years.

Study Sample: 463 families eligible for and interested in participating in Even Start family literacy services.

Intervention: Even Start families were offered family literacy services, defined as (1) interactive parent-child literacy activities, (2) parenting education, (3) adult education, and (4) early childhood education.

Research Design: Randomized controlled field trial in which families were randomly assigned either to Even Start (309 families) or a control group (154 families).

Control or Comparison Condition: Control families could participate in any educational and social services to which they were entitled, but they were not allowed to participate in Even Start for one year.

Data Collection and Analysis: Pretest data on child and adult literacy skills were collected in the fall, posttest data were collected in the spring/summer, and follow-up data were collected the next spring. Measures included direct assessment of children (Peabody Picture Vocabulary Test, Woodcock-Johnson Battery, Story and Print Concepts), direct assessment of parents (Woodcock-Johnson Battery), teacher report on children (Social Skills Rating System), parent reports on economic and educational status, child literacy-related skills, home literacy environment and activities, parent assessment of children (Vineland Communication Domain), and school records. A longitudinal sample (data at all three waves) of children and parents was created for each outcome measure, and t-tests were conducted to assess differences in gains between Even Start and control groups. The sample size for the analysis of any given outcome depends on several factors including attrition, age of the child, exclusion of families who were assessed in Spanish, and the need for longitudinal data. For example, the PPVT analysis for children was done with samples of 97 Even Start and 44 control children, and the

Woodcock-Johnson analysis for parents was done with samples of 149 Even Start and 65 control parents.

Findings: As was the case at posttest, Even Start children and parents made gains on a variety of literacy assessments and other measures at follow-up, but they did not gain more than children and parents in the control group. It had been hypothesized that follow-up data might show positive effects because (1) Even Start families had the opportunity to participate for a second school year, and (2) change in some outcomes might require more time than others. However, the follow-up data do not support either of these hypotheses.

Conclusion: The underlying premise of Even Start as described by the statute and implemented in the field was not supported by this study.

Notes

[1] Our suggested list of key items is drawn, to a large extent, from the following authoritative sources: Evidence Standards of the Institute of Education Sciences' What Works Clearinghouse, at www.w-w-c.org/reviewprocess/standards.html; the Standards of Evidence of the Society for Prevention Research, in Brian R. Flay et. al., "Standards of Evidence: Criteria for Efficacy, Effectiveness, and Dissemination," Prevention Science, forthcoming, September 2005 (available online at http://www.preventionresearch.org/commlmon.php#SofE); the Consolidated Standards of Reporting Trials (CONSORT) Statement, in David Moher, Kenneth F. Schulz, and Douglas Altman, "The CONSORT Statement: Revised Recommendations for Improving the Quality of Reports of Parallel-Group Randomized Trials," JAMA, vol. 285, no. 15, April 18, 2005, pages 1987-1991.

[2] The structured abstract format suggested by the Institute of Education Sciences is based on the concept proposed in Frederick Mosteller, Bill Nave, and Edward J. Miech, "Why We Need a Structured Abstract in Education Research," Education Researcher, vol. 33, no. 1, January/February 2004, pp. 29-34.

[3] Preferably, you should also indicate the extent to which those eligible for the study, and those invited or selected, actually became members of the study sample.

[4] If a power analysis is undertaken, you should report not only its results, but also its statistical assumptions (e.g., desired power, minimum detectable effect size, intra-class correlation coefficient, and fixed or random effects).

[5] We suggest you report here only the budgetary cost of the intervention (e.g., in the tutoring example, the cost of program materials, and paying the tutors and other program staff). If the intervention increases or decreases the cost of other services (e.g., reduces the number of students referred to special education classes), these effects should be reported in the Results section of the report (section 4).